DOCUMENT RESUME

ED 049 275 TM 000 447

AUTHOR	Whalen, Thomas E.
TITLE	A Validation of the Smith Test for Measuring Teacher Judgment of Written Composition.
PUB DATE	Feb 71
NOTE	7p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	*Composition Skills (Literary), *Essay Tests, *Reliability, *Scoring, *Test Validity, Writing Skills

ABSTRACT

Smith (1969) reported the results of an instrument for measuring teacher judgment of written composition. His test was first administered to a group of "experts" whose ratings were in high agreement. Then the test was given to a sample of over 200 teachers and lay readers. Among Smith's conclusions was that over half of the teachers have judgment which differs significantly from the experts. This study sought to determine if rater differences as measured by Smith's test would remain constant for another set of essays. Six raters were selected, on the basis of their scores on Forms A and B of the Smith test, to read and score 71 seventh-grade essays. No significant differences were observed between "good" and "bad" raters. The results cast doubt on the validity of Smith's test as a general instrument for assessing essay-rating behavior. Although the test does appear to separate raters in terms of their rankings of essays, and even though these rankings are relatively reliable, difference between raters did not remain constant for another set of essays. (Author/LR)

A VALIDATION OF THE SMITH TEST FOR MEASURING TEACHER JUDGMENT OF WRITTEN COMPOSITION

Thomas E. Whalen

Literature on the measurement of writing ability is replete with evidence of the unreliability and/or invalidity of reader evaluation of student writing. Schumann (1968) stated that "research indicates that the youngster who has neat penmanship will get at least a "C" grade in composition work irrespective of what he actually says" (p. 1163).

As early as 1921, Hopkins demonstrated that the score a student made on a College Board examination might well depend more on which year he appeared for the examination, or on which person read his paper, than it would on what he had written. Godshalk and others (1966) presented a definitive review of the shifts in College Board testing procedure from its inception. They concluded that the two main sources of unreliability were (1) differences in quality of student writing from one topic to another, and (2) the differences among readers in what they consider the characteristics of good writing.

Evidence to support the second source of unreliability above was presented by Diederich and French (1961). The authors conducted a factor analytic study involving fifty-three readers from six different professional areas. The study revealed five "schools of thought" with regard to measuring composition skill: (1) ideas, (2) form, (3) flavor, (4) mechanics, and (5) wording.

The five reader-factors were identified by a "blind" classification of 11,018 comments written on 3,557 papers. The readers included college English teachers, social scientists, writers and editors, lawyers, natural scientists, and business executives. Ninety-four percent of the papers received seven or more of the nine possible grades, and no paper received less than five different grades.

The median correlation between readers was .31. Readers in each field agreed slightly better with the English teachers than with one another. Three College Board tests taken by the student writers formed a separate test-factor that had practically zero correlation with all reader-factors except mechanics (.50) and wording (.45).

Despite this apparently overwhelming evidence of rater inconsistency, efforts continue to be made, and rightly so, toward the achievement of more reliable procedures for assessing students' writing. Smith (1969) reported the results of an instrument for measuring teacher judgment in the evaluation of written composition. He constructed a test to determine how well teachers agree in their rating behavior with a set of expert English teachers. The test consists of two forms, A and B, each containing five short essays taken from the Sequential Tests of Educational Progress, Essay Test, and from other samples of actual student writing. Raters are asked simply to rank the five essays on each form from best to worst.

Smith first administered the test to a group of five "experts." The raters in this group were all secondary English teachers "who had been formally recognized as outstanding in the teaching of composition within their school districts or by some outside agency" (p. 187). Impressive reliability coefficients were reported for the expert raters. Inter-rater reliabilities (using Snedecor's formula) ranged from .840 to .920 for two administrations for forms A and B. Reliabilities of average ratings ranged from .963 to .983. The test-retest reliability was reported as 1.00 (p.188).

The test was then administered to a sample of over 200 teachers and lay readers to determine the extent of their agreement with the experts. Smith found much greater variance among subjects in the sample population than among the experts. Among the conclusions reached by Smith were the following:

1. Judgment as measured by this test is not related to experience, academic background or professional training.

2. More than half the teachers disagree to some extent with the experts in judgment as measured by this test.

3. Between ten and twenty percent of classroom teachers have judgment that is contrary to that of the experts, and thus, "these persons are not competent to make such judgments" (p. 190).

As possible applications for his test, Smith suggests its use as part of a battery of tests to screen composition reader applicants, and as a tool to screen raters in research when judgment in the evaluation of written composition is a factor. It might also be used to provide individual and prospective teachers with knowledge of their judgment in the evaluation of written composition (p. 193).

The purpose of the present study was to determine to what extent the results of Smith's test can be generalized to other essay-rating situations. If Smith's test can, indeed, provide valid measurements of rating behavior, then differences between raters on his test should remain constant acorss other samples of writing judged by the same raters.

METHOD

Forms A and B of Smith's test were administered to thirty-three individuals including nineteen elementary and secondary teachers and fourteen graduate students in educational psychology. Scores from the two forms were combined (a procedure suggested by Smith to increase reliability), producing a scale from zero to ten. High scores (8, 9, or 10) indicated agreement with the experts; low scores (0 through 5) represented disagreement; scores of 6 and 7 indicated judgment that is borderline.

Six of the thirty-three raters were selected on the basis of their test scores to read and score seventy-one seventh grade essays. The essays were gathered from three seventh grade English classes of average ability. All students wrote on the same topic--their reactions to the novel The Adventures of Tom Sawyer. The essays were all approximately 200 words in length. None of the raters was associated with the school from which the essays were selected nor had any knowledge of the students whose essays he rated.

Four of the six raters were in high agreement with the experts on Smith's test. They achieved scores of 8,8,9, and 10 on the test. Two of the raters were in complete disagreement, having both received scores of four. A reliability coefficient for average ratings (Ebel, 1951) was calculated for the group of four who were in agreement. A second group of raters was formed which included the two raters in disagreement with the experts and two who were in high agreement (randomly selected from the previous group of four). The interjudge reliability was calculated for the second group and was compared statistically with the coefficient for the group in complete agreement. In addition, an intercorrelation matrix of all six judges' ratings was generated. Coefficients in this matrix were compared to determine the extent of agreement between individual raters.

RESULTS AND DISCUSSION

For the sample of thirty-three teachers and graduate students, the scoring range on Smith's test was from three through ten. The mean score was 6.27 with a standard deviation of 1.91. Fully two-thirds of the sample disagreed to some extent with the experts (scores of seven or below). One-third of the raters were in complete disagreement (a score of five or less). A comparison of mean scores for teachers versus graduate students showed no appreciable difference between groups -- 6.36 and 6.21, respectively. In general, these findings were in accord with the results of Smith's research except that a somewhat greater percentage of persons in this sample had judgment contrary to that of the experts.

The results of the analysis of variance to determine the reliability of averaged scores for both groups is given in Table 1. This analysis is appropriate when the raters' scores are eventually averaged and is designed to eliminate variance due to the raters' operating at different means. In this study, the readers were asked to rate each of the seventy-one essays by assigning a grade of "A", "B", "C", or "D". These ratings were then quantified on a 4-point scale.

The reliability for the four judges in agreement was .79. For the mixed group composed of two judges in agreement and two in disagreement, the coefficient of reliability was .84. Thus, a higher reliability was calculated for the group whose members had demonstrated opposing views with regard to the essays on Smith's test. These two coefficients were compared (Lordahl, 1967) and were found not to differ significantly.

Table 2 shows the correlation matrix for all six judges' scores. Judges 5 and 6 were the two in disagreement with the experts. Judges 2 and 4 were those selected for inclusion in Group II. A comparison of coefficients in the matrix indicated that Judge 5, who scored low on the Smith test, was in high agreement with Judges 2 and 4, who scored high on the test. In fact, the average correlation (using Fisher's z-transformation) between this low-scoring judge and the two high-scoring judges was greater than the correlation between the two high-scoring judges themselves. Judge 6 disagreed to a greater extent with Judges 2 and 4. However, his average correlation with the high-scoring judges indicated that he also agreed with them to a greater extent than they agreed with one another.

The evidence in this investigation casts doubt on the validity of Smith's test as a general instrument for assessing essay-rating behavior. Although the test does appear to separate raters in terms of their rankings of the ten essays, and even though these rankings are relatively reliable measures (.87 for test-retest using combined scores from forms A and B), differences between raters did remain constant for another set of essays judged by the same raters. Additional research is necessary before this test should be applied seriously to any of the

TABLE 1

ANALYSIS OF VARIANCE RELIABILITIES OF AVERAGED RATINGS FOR ESSAY GRADES

GROUP I
(In Agreement)

Source	DF	SS	MS
Essays	70	132.92	1.90
Raters	3	4.42	1.47
Error	210	82.58	0.39
Total	283	219.92	

Reliability = 1-(MS error/MS essays) = 0.79

GROUP II
(In Disagreement)

Source	DF	SS	MS
Essays	70	149.77	2.14
Raters	3	3.20	1.07
Error	210	73.80	0.35
Total	283	226.77	

Reliability = 1- (MS error/MS essays) = 0.84

TABLE 2

Intercorrelations of Six Raters

Rater	1	2	3	4	5	6
1	---	.44	.51	.50	.51	.39
2		---	.46	.52	.65	.67
3			---	.52	.60	.42
4					.63	.40
5					---	.52
6						---

REFERENCES

Diederich, P., French, J., and Carlton, S. Factors in the judgment of writing ability. ETS Research Bulletin, No. 15. Princeton: Educational Testing Service, 1961.

Ebel, R. L. Estimation of reliability ratings. Psychometrika, 1951, 16 (4) 407-424.

Godshalk, F., Swineford, F., and Coffman, W. The measurement of writing ability. Princeton: Educational Testing Service, 1966.

Hopkins, L.T. The marking system of the College Entrance Examination Board. Harvard Monographs in Education, Series 1, No. 2, Cambridge, Mass.: The Graduate School of Education, Harvard University, October 1921.

Lordahl, D. S. Modern Statistics for behavioral sciences. New York: The Ronald Press Co., 1967.

Schumann, P.F. What criteria do you use for grading compositions? English Journal, 1968, 57 (8), 1163-1165.

Smith, V. H. Measuring teacher judgment in the evaluation of written composition. Research in the Teaching of English, 1969, 3 (2), 181-195.